Aplikasi untuk Membangun Corpus dari Data Hasil *Crawling* dengan Berbagai Format Data Secara Otomatis

Jati Sasongko
Program Studi Teknik Informatika
Fakultas Teknologi Informasi, Universitas Stikubank
jati@unisbank.ac.id

Abstrak

Aplikasi membangun *corpus* dengan berbagai format data dibangun dari beberapa proses atau modul sehingga membentuk aplikasi yang befungsi untuk membangun sebuah *corpus*. Proses-proses yang terdapat pada aplikasi membangun *corpus* terdiri dari : pengumpulan data file teks dan image (*crawling*), poses pencarian file teks dan image dalam folder atau direktori (*searching*), proses *input* data file teks dan image ke database (*corpus*), proses pengindeksan (*indexing*).

Aplikasi mampu menampilkan hasil pencarian dokumen dan mengurutkannya berdasarkan urutan dari penemuan dari file data yang dicari, dalam arti dokumen data yang ditemukan pertama kali akan ditempatkan diurutan pertama sedangkan dokumen data yang ditemukan terakhir akan ditempatkan pada urutan paling bawah. Aplikasi juga mampu melakukan konversi dari dokumen teks dengan berbagai format data ke dalam bentuk dokumen teks txt, juga dalam melakukan konversi pada semua format file image ke dalam bentuk format bmp. Konversi dilakukan untuk menyamakan format untuk dapat mempermudah dalam penyimpanan dalam database.

Kata Kunci : *Corpus, Crawling, Retrieval Information*

PENDAHULUAN

1. Latar Belakang

Membangun corpora dengan skala yang besar untuk digunakan dalam analisis linguistik dalam bentuk digital dapat dianggap sebagai hasil dari suatu kumpulan rujukan, dikumpulkan dan disusun dalam bentuk tulisan tangan selama puluhan tahun. Namun, istilah korpus saat ini yang paling banyak digunakan untuk merujuk kepada sekumpulan data linguistik yang dikumpulkan untuk tujuan analitik tertentu, dengan anggapan bahwa hal itu akan disimpan, dikelola, dan dianalisa dalam bentuk digital. Bapak korpus linguistik jenis ini adalah Brown Corpus, dibuat di Brown University di awal tahun enam puluhan, menggunakan metode yang masih relevan hingga sekarang. Ahli bahasa dan linguistik telah berkembang dari berbagai pendekatan berbasis korpus untuk subyek disiplin akademis yang diakui telah mempunyai andil. Namun demikian, linguistic berbasis korpus secara luas masih dianggap sebagai pusat penelitian dengan banyak aspek dalam sifat dan fungsi bahasa manusia, dengan aplikasi di bidang-bidang yang beragam seperti leksikografi, pemrosesan bahasa alamiah, mesin terjemahan, dan belajar bahasa. Bahwa dalam pembangunan corpus dengan menggunakan banyak format data banyak sekali kendala yang dihadapi seperti pembuatan corpus dalam konsorsium BNC. Dalam penelitian ini data yang digunakan berasal dari hasil mesin crawler yang dibagi dalam dua format data, yang pertama, file teks dengan menggunkan bahasa apapun dan yang kedua menggunakan file image.

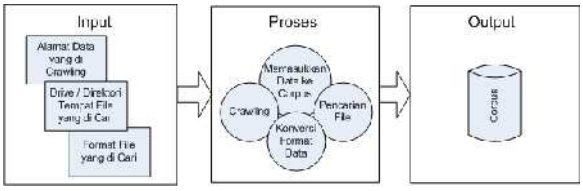
2. Tujuan

Penelitian ini dilakukan dengan tujuan menghasilkan suatu aplikasi yang dapat digunakan untuk membangun corpus dengan berbagai format data yang berasal dari hasil crawling secara otomatis.

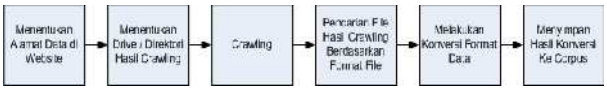
3. Metode Penelitian

Analisa dan Induksi yang akan digunakan dalam mencapai tujuan

Perolehan data diambil dari alamat sebuah alamat website di internet dengan menggunakan crawler. Yang perlu dilakukan dalam mesin crawler ini yaitu pertama, menentukan alamat website dimana tempat data tersebut berada, kedua, menentukan kedalaman atau berapa dalam level dalam mengambil data ini dilakukan, ketiga, menentukan apakah dalam mengambil data ini hanya dari server yang sama atau boleh dari server yang lain, keempat, menentukan tempat penyimpanan data hasil crawling ini. Setelah menentukan beberapa hal yang telah disebutkan sebelumnya baru dilakukan proses pengambilan data, proses ini akan dilakukan terus menerus sampai data yang terdapat dalam link tersebut habis, atau proses pengambilan data dibatalkan. Data yang diambil dari proses crawling ini akan disimpan dalam direktori atau drive yang telah ditentukan sebelumnya. Data yang tersimpan dalam bentuk file dengan berbagai format data, untuk data teks, ada yang berformat doc, txt, htm, html, dsb, sedangkan untuk data image, ada yang berformat jpg, png, gif, dsb. Data yang telah terkumpul berikutnya diproses untuk dapat dijadikan sebuah corpus dalam bentuk tabel atau database. Dalam hal ini yang pertama kali dilakukan yaitu pencarian data yang telah tersimpan di dalam folder atau drive, pencarian dilakukan dengan menentukan tipe data yang sesuai dengan kebutuhan yang akan dimasukkan ke dalam database corpus, tipe data yang dimaksud dapat berupa file doc, txt, html, jpg, png, gif dsb. Hasil dari pencarian akan ditampilkan secara urut berdasarkan urutan pencarian. Data dari hasil pencarian tersebut secara urut pula akan dimasukkan dalam database corpus. Data yang dimasukkan dalam database corpus dibagi menjadi tiga bagian : id, url dan isi. Pada saat memasukkan data dalam database corpus dilakukan sub-sub proses yang gunanya untuk menyamakan tipe data yang berbeda sehingga dari berbagai format data yang ada dapat dimasukkan dalam database corpus. Hal inilah yang menjadi kunci utama dalam penelitian ini, sehingga dari berbagai format data yang ada dapat dibangun menjadi corpus.



Gambar 1. Diagram Rancangan Penelitian



Gambar 2. RoadMap Rancangan Penelitian

TINJAUAN PUSTAKA

1. *Corpus*

Penelitian empiris dapat dilakukan dengan menggunakan teks tertulis atau lisan, seperti teks-teks dasar dari berbagai jenis sastra dan analisis linguistik. Tapi gagasan tentang korpus sebagai dasar untuk sebuah bentuk linguistik empiris berbeda dalam beberapa cara mendasar dari teks-teks tertentu. Pada prinsipnya, setiap koleksi lebih dari satu teks dapat disebut corpus: istilah *corpus* dalam bahasa latin berarti *body*, maka corpus dapat didefinisikan sebagai isi setiap teks. Tapi istilah 'corpus' ketika digunakan dalam konteks linguistik modern memiliki konotasi yang lebih spesifik. Ada empat karakteristik dari corpus (McEnery dan Wilson, 2001) :

a. Sampling and Representativeness

Dalam membangun sebuah korpus dari berbagai bahasa, dapat ditarik dari sebuah sampel yang mewakili dari berbagai pengujian secara maksimal, yaitu menyediakan corpus seakurat mungkin dari kecenderungan yang beragam *term*asuk proporsi antara corpus dan informasi yang dicari. Jadi, tidak semata-mata berdasarkan pada teks sampel yang dipilih, akan tetapi mencari sampel dari berbagai sumber yang diambil dari sumber dokumen aslinya, sehingga akan memberikan gambaran yang cukup akurat dari seluruh informasi yang akan didapatkan.

b. Finite Size

Selain sampling, istilah corpus juga cenderung menyiratkan suatu isi teks dengan ukuran yang terbatas, misalnya 1.000.000 kata. Teks dapat terus ditambahkan ke dalamnya, sehingga semakin besar karena lebih banyak

sampel yang ditambahkan. Keuntungan utamanya : (1) teks menjadi tidak statis karena teks yang baru akan selalu ditambahkan dan (2) ruang lingkup akan lebih besar dan jauh lebih luas sehingga akan mencakup dari bahasa yang digunakan. Kelemahan utamanya adalah bahwa, karena terus berubah dalam ukuran dan kurang ketatnya sampel, menjadi sumber yang kurang terpercaya dalam segi kuantitatif (sebagai lawan kualitatif). Jadi sebaiknya pada awal pembangunan korpus, rencana riset ditetapkan secara rinci bagaimana berbagai bahasa yang digunakan diambil sampelnya, berapa banyak sampel dan kata harus dikumpulkan sehingga jumlah keseluruhan yang sudah ditetapkan ini dapat digunakan.

c. Machine-Readable Form

Corpora yang dapat dibaca oleh mesin memiliki beberapa keunggulan dibandingkan dengan format tertulis atau lisan. Pertama dan paling penting keuntungan dari corpora yang dapat dibaca oleh mesin adalah bahwa dimungkinkan untuk mencari dan memanipulasi dengan cara-cara yang tidak dilakukan dengan format lain. Sebagai contoh, sebuah korpus dalam format buku, akan perlu dibaca dari depan sampai belakang untuk mengambil semua contoh kata, dengan korpus yang dapat dibaca oleh mesin, tugas ini dapat dicapai dalam beberapa menit dengan menggunakan perangkat lunak, atau sedikit lebih lambat, dengan menggunakan fasilitas pencarian di pengolah kata. Keuntungan kedua corpora yang dapat dibaca oleh mesih adalah bahwa dapat dengan cepat dan mudah diperkaya dengan informasi tambahan.

d. Standard Reference

Meskipun tidak *term*asuk hal yang penting dari definisi suatu korpus, tetapi ada juga pemahaman bahwa korpus merupakan referensi standar untuk berbagai bahasa yang diwakilinya. Hal ini mengandaikan ketersediaan yang luas kepada peneliti lain, keuntungan dari korpus yang tersedia secara luas adalah bahwa akan memberikan tolok ukur yang dapat digunakan sebagai pembanding dalam studi. Misalnya. secara langsung dibandingkan dengan hasil yang dipublikasikan (selama metodologi sama) tanpa perlu perhitungan ulang. Korpus standar juga berarti penggunaan corpus yang sama digunakan untuk berbagai macam variasi studi dan yang membedakannya yaitu penggunaan data pengujiannya dan metodologi yang digunakan dalam studi. (McEnery dan Wilson, 2001).

2. *Crawler*

Crawler adalah proses untuk mengumpulkan informasi dari halaman web berdasarkan indeks. Tujuan dari *crawler* adalah dengan cepat dan efisien mengumpulkan banyak informasi dari halaman web yang berguna, berikut dengan struktur link yang terkoneksi dengan halaman web tersebut.

Fitur *crawler* yang perlu disediakan :

a. Didistribusikan: *crawler* harus memiliki kemampuan untuk dapat dijalankan dalam berbagai macam model hardware dan software yang berbeda.

b. *Scalable*: arsitektur *crawler* dimungkinkan untuk dapat meningkatkan kemampuan dengan menambahkan hardware dan *bandwith*.

c. Kinerja dan efisiensi: *crawler* dapat membuat lebih efisien dari penggunaan berbagai sumber daya sistem, termasuk prosesor, penyimpanan, dan bandwidth jaringan.

d. Kualitas: *crawler* dapat melayani kebutuhan permintaan anggota dengan mengambil yang berguna dari halaman web yang diketemukan.

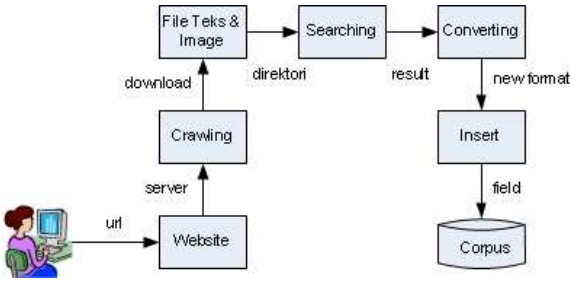
e. *Freshness*: *crawler* harus dapat mendapatkan informasi dari halaman web yang belum dikunjungi dan tidak mengambil informasi yang sudah pernah diambil.

f. *Extensible*: *crawler*s harus dirancang untuk dapat disesuaikan dengan perkembangan teknologi yang ada sekarang dan yang akan datang, baik format data, protocol yang digunakan, dan seterusnya. (Pinkerton, 2000).

PERANCANGAN

1. *Logical Process* Aplikasi dalam Membangun Corpus

Aplikasi membangun *corpus* dengan berbagai format data dibangun dari beberapa proses atau modul sehingga membentuk aplikasi yang befungsi untuk membangun sebuah *corpus*. Proses-proses yang terdapat pada aplikasi membangun *corpus* terdiri dari : pengumpulan data file teks dan image (*crawling*), poses pencarian file teks dan image dalam folder atau direktori (*searching*), proses *input* data file teks dan image ke database (*corpus*), proses pengindeksan (*indexing*). *Logical process* aplikasi membangun corpus secara diagram dapat dilihat dalam gambar 3.



Gambar 3. Diagram *Logical Process* Aplikasi Membangun Corpus

Masing-masing proses sistem temu kembali informasi dapat dijelaskan sebagai berikut :

a. Pengumpulan Data Dokumen menggunakan *Crawler*

Crawler merupakan program yang berjalan secara otomatis, berisi *script* program yang melakukan *crawling* melalui halaman website untuk mengumpulkan data berdasarkan indeks dari halaman web yang ditemukan. Nama alternatif untuk *crawler* seperti *spider*, *robot* dan *automatic indexer*. Sebuah *crawler* dapat digunakan untuk berbagai tujuan. Penggunaan paling umum terkait dengan mesin pencari. Mesin pencari menggunakan *crawler* untuk mengumpulkan informasi yang terdapat pada halaman website sehingga ketika pengguna internet memasukkan kata kunci pencarian dapat dengan cepat memberikan informasi yang relevan kepada user.

Dalam penelitian ini *crawler* digunakan untuk mengumpulkan data file teks dan image yang diperolah dari sebuah alamat website yaitu www.unisbank.ac.id Ada beberapa sub proses dalam proses pengumpulan data dengan menggunakan *crawler* ini. Dimulai dari menentukan alamat web yang akan diambil datanya, kemudian membuat koneksi terhadap alamat yang telah ditentukan sebelumnya, apabila koneksi gagal akan diberikan informasi bahwa koneksi tidak dapat dilakukan, apabila koneksi berhasil maka akan dilanjutkan proses selanjutnya yaitu mendownload informasi yang ada di alamat web tersebut.

Proses pengunduhan akan dilakukan sampai seluruh informasi yang ada di alamat website tersebut sudah tidak ada lagi yang diambil, termasuk link-link indeks yang terhubung dengan alamat web tersebut. Pengunduhan informasi dapat berupa file-file gambar, html, pdf, dan sebagainya. Hasil pengunduhan akan disimpan secara otomatis ke dalam folder yang juga telah ditentukan sebelumnya. File-file yang terkumpul kemudian dipilih untuk mendapatkan file dokumen teks dan image yang akan digunakan untuk dimasukkan di dalam tabel *corpus* yaitu berupa file txt dan bmp.

b. Memasukkan Dokumen Teks ke dalam Tabel *Corpus*

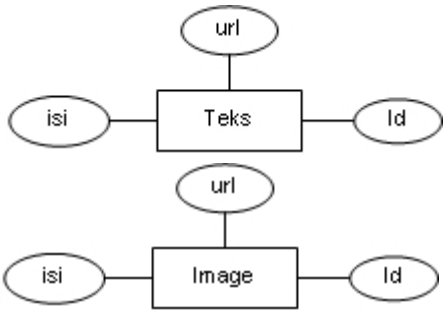
Langkah awal dari sistem ini untuk dapat berfungsi dengan semestinya dimulai dari memasukkan data dokumen yang telah diunduh melalui *crawler* ke dalam tabel *corpus*. Tahapan prosesnya yaitu menjalankan modul pendataan file teks dan image, dimana modul ini akan digunakan untuk melakukan pencarian data file teks dan image sesuai dengan kriteria. Dalam penelitian ini kriteria yang dimaksud yaitu pencarian file dengan format teks txt dan bmp. Pencarian dilakukan dengan menentukan direktori termasuk subdirektorinya tempat file tersebut berada. Ketika file-file tersebut ditemukan maka hasilnya akan ditampilkan berserta nama subdirektorinya. Selanjutnya dilakukan proses memasukkan hasil pencarian file yang dicari ke dalam tabel *corpus*. Saat melakukan proses tersebut ditampilkan isi dokumen, nama file atau pathnya. Saat menyimpan file teks dan image juga dilakukan proses pemberian kode terhadap file data

tersebut. Sehingga struktur data yang tersimpan di dalam tabel *corpus* terdiri dari kode, isi dan nama file atau path file dokumen.

Proses memasukkan dokumen file html ke tabel *corpus* akan membutuhkan beberapa waktu tergantung dari banyaknya hasil pencarian yang dilakukan sebelumnya. Apabila hasil pencariannya banyak maka waktu yang dibutuhkan akan lebih lama, sebaliknya apabila hasil pencarian file yang dilakukan menghasilkan jumlah file yang relatif sedikit maka waktu yang dibutuhkan lebih pendek. Dan pada saat melakukan proses memasukkan file teks dan image ke tabel *corpus* terkadang akan berhenti yang dikarenakan adanya file yang kosong sehingga tidak ada data yang dapat dimasukkan ke dalam tabel *corpus*, dan apabila berhenti maka file tersebut harus dihapus dari daftar hasil pencarian file dan dilanjutkan kembali.

2. *Entity Relationship Diagram*

Dalam pemrosesan data, metode pemodelan data menggunakan ERD (*Entity Relationshyp Diagram*) atau Diagram Hubungan Entitas yang memungkinkan perekayasaan perangkat lunak untuk mengidentifikasi objek data dan hubungannya dengan menggunakan notasi grafis.



Gambar 4. Entity Relationship Diagram Aplikasi Membangun Corpus

Entity Relationshyp Diagram digunakan untuk memudahkan struktur data dan hubungan antar data, karena hal ini relatif kompleks. Dengan *Entity Relationship Diagram* dapat melakukan pengujian model dengan mengabaikan proses yang harus dilakukan. Dalam rancangan sistem basis data untuk Sistem Temu Kembali Informasi Bahasa Inggris, digunakan *Entity Relationshyp Diagram* atau Diagram Hubungan Entitas dan desain tabel untuk menggambarkan atribut-atributnya yang ditunjukkan pada Gambar 4.

3. Perancangan Database

Dari proses (modul) aplikasi membangun *corpus* akan terdapat 2 tabel yang terdiri dari tabel teks dan tabel image di dalam database *corpus*.

a. Rancangan Tabel Teks

Tabel teks digunakan untuk menyimpan data teks. Sistem akan melakukan proses mengambil file teks dari folder atau direktori kemudian menyimpannya dalam tabel. Dari proses ini data yang akan disimpan berupa kode, isi dan nama file dari data yang diproses. Struktur tabel teks dapat dilihat seperti pada tabel 1.

Tabel 1. Rancangan Tabel Teks

Field	Type	Collation	Attributes	Null	Default	Extra
kode	varchar(25)	utf8_general_ci		No		
url	text	utf8_general_ci		No		
isi	longblob		BINARY	No		

b. Rancangan Tabel Image

Tabel image digunakan untuk menyimpan data image. Aplikasi akan melakukan proses mengambil file image dari folder atau direktori kemudian menyimpannya dalam tabel. Dari proses ini data yang akan disimpan berupa kode, isi dan nama file dari data yang diproses. Struktur tabel image dapat dilihat seperti pada tabel 2.

Tabel 2. Rancangan Tabel Image

Field	Type	Collation	Attributes	Null	Default	Extra
kode	varchar(25)	utf8_general_ci		No		
url	text	utf8_general_ci		No		
isi	longblob		BINARY	No		

4. Diagram Alir Data

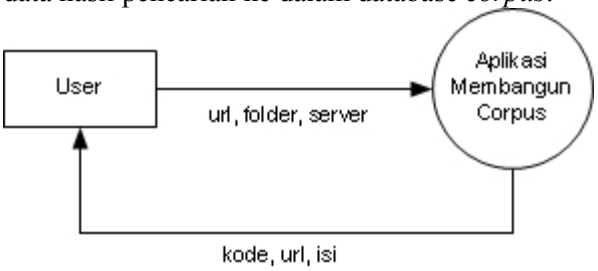
Diagram konteks aplikasi membangun corpus memiliki dua entitas luar yaitu image dan teks. Sistem temu kembali informasi menerima *input query* dalam teks bahasa indonesia. Dan sistem akan mengeluarkan *output* yaitu dokumen yang dicari dengan *query* yang di*input*kan oleh user.

Diagram alir dokumen level 1 sistem temu kembali informasi terdiri dari proses *input* data

dokumen, proses *input* data kamus indonesia-inggris, proses *input* data *stopword*, proses *input* data *punctuation*, proses tokenisasi *(tokenization*), proses penghilangan *stopword* (*stopword removal*), proses penghilangan tanda baca (*punctuation removal*), proses *term indexing*, proses proses terjemah (*translating*), proses *query* dan proses *ranking*.

a. Diagram Alir Data Level 0

Diagram level 0 pada gambar 5. menggambarkan proses yang terdapat dalam aplikasi membangun *corpus*. Dimana user dapat melakukan pengumpulan data secara otomatis dari sumber data melalui jaringan internet, pengumpulan data dilakukan dengan menggunakan mesin crawler dengan user harus memasukkan alamat internet yang dituju, server yang akan didownload dan folder tempat data hasil download akan disimpan. Proses yang lain yaitu melakukan pencarian dengan kriteria tertentu dari data yang telah terkumpul dari folder yang telah ditentukan dan memasukkan data hasil pencarian ke dalam database *corpus*.

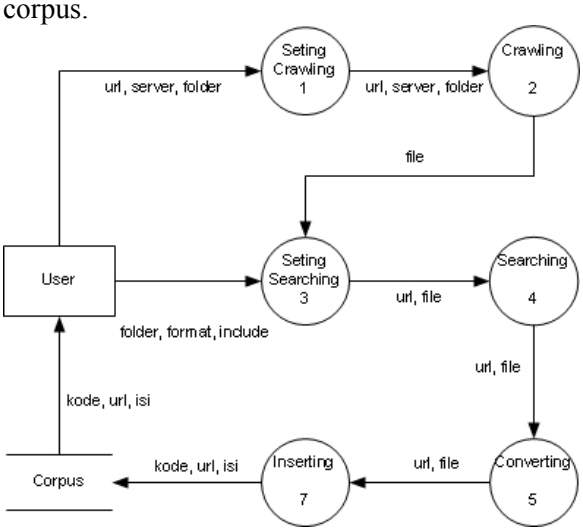


Gambar 5. Diagram Alir Data Level 0 Aplikasi Membangun Corpus

b. Diagram Alir Data Level 1

Diagram Alir Data Level 1 pada gambar 6 menggambarkan proses-proses yang terdapat dalam aplikasi dalam membangun sebuah *corpus*. Ada proses-proses yang terjadi berhubungan langsung dengan user atau dengan kata lain yang melakukan proses tersebut atas perintah dari user seperti menentukan url, server dan folder, tetapi ada juga proses-proses yang berjalan secara otomatis dalam arti bahwa proses-proses tesebut akan berjalan secara otomatis apabila mendapatkan masukkan informasi atau data dari proses sebelumnya. Secara berurutan proses-proses yang terdapat dalam diagram alir data level 1 ini dimulai dari proses seting crawling yaitu proses untuk memasukkan url dari suatu alamat website tempat data teks maupun image yang akan didownload sebagai data primer. Selain menentukan url juga menentukan berapa tingkatan link data di dalam website tersebut yang akan didownload, tingkatan ini sangat mempengaruhi dari banyaknya data yang akan didownload, semakin tinggi tingkatannya maka semakin banyak data yang akan didonwload. Berikutnya menentukan server yang akan didownload, apakah data didownload dari hanya satu server atau lebih, apabila memilih satu server maka apabila terdapat link data yang berada pada server yang berbeda tidak akan didownload, tetapi sebaliknya apabila memilih lebih dari satu server maka apabila ada link data yang terdapat pada server berbeda dan masih dalam seting yang ditentukan maka data tersebut akan didownload. Setelah seting crawling dilakukan semuanya maka tinggal melakukan proses crawling yaitu proses untuk melakukan download data dari url yang telah ditentukan, seberapa banyak data yang akan download berdasarkan dari url dan seting tingkatan level link yang telah ditentukan sebelumnya, sehingga data yang termasuk dalam setingan tersebut akan didownload semuanya. Data hasil download dari mesin robot tersebut akan disimpan dalam sebuah folder yang juga telah ditentukan sebelumnya sebelum proses crawling dilakukan. Setelah semua data terdownload berikutnya melakukan proses seting searching, proses ini dilakukan untuk menentukan folder dimana data hasil download berada atau mungkin juga menentukan folder dari data yang sudah dimiliki dari sumber lain. Selain menentukan folder juga menentukan format data yang akan dicari, misalnya yang dicari format data *.txt atau yang lain, dengan menentukan format data yang dicari maka akan mempercepat proses pencarian juga akan mempermudah dalam mengolah data hasil dari pencarian untuk dapat dilakukan proses berikutnya. Berikutnya juga menentukan apakah data yang dicari berada pada satu folder atau termasuk dalam subfolder, hal ini akan sangat mempengaruhi dari hasil pencarian, apabila memilih dengan subfolder maka hasil pencarian akan menghasilkan data lebih baik dari pada hanya memilih satu folder tanpa subfolder. Setelah seting searching dilakukan maka proses searching dapat dilakukan dan hasilnya akan ditampilkan berdasarkan urutan proses

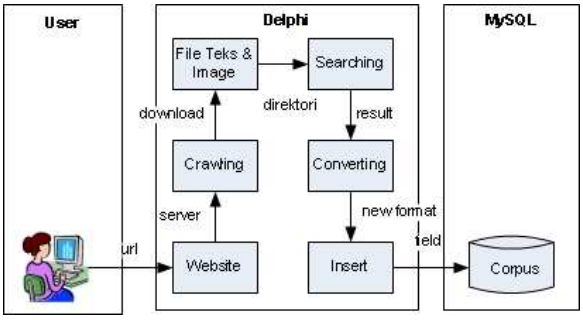
pencarian, dalam arti bahwa data yang ditemukan pertama akan ditempatkan pada urutan teratas dan sebaliknya data yang ditemukan terakhir maka akan ditempatkan pada urutan terakhir. Data yang dihasilkan dari proses pencarian berikutnya akan dilakukan proses converting yaitu proses untuk melakukan perubahan format data dari format asli data yang diambil dari proses crawling diubah menjadi format data untuk dapat masuk ke dalam database corpus, dalam hal ini semua file teks akan diubah dalam bentuk txt sedangkan semua file image akan diubah ke dalam format bmp. Setelah semua file data dikonversi maka tinggal dimasukkan dalam database corpus, pada saat memasukkan dalam database corpus field yang dimasukkan kode, url dan isi. Proses insert ini merupakan proses terakhir dari seluruh proses yang terdapat dalam aplikasi membangun corpus.



Gambar 6. Diagram Alir Data Level 1 Aplikasi Membangun *Corpus*

5. Implementasi *Logical Process* Sistem Pencarian Dokumen

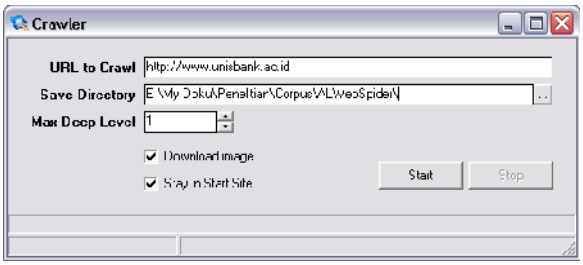
Spesifikasi hardware dan perangkat lunak yang digunakan untuk merancang dan membuat sistem ini menggunakan processor intel core 2 duo T5200 @1,6 GHz, memori 1 GHz. Sistem operasi menggunakan Windows XP, penyimpanan dan pemrosesan data menggunakan Mysql, user interface dan bahasa pemrograman yang digunakan menggunakan Borland Delphi. Koneksi tabel menggunakan ODBC. Gambar 7. menunjukkan implementasi perangkat lunak untuk masing masing proses.



Gambar 7. Implementasi *Logical Process* Aplikasi Membangun *Corpus*

6. Implementasi Seting Crawler

Seting Crawler dalam implementasi aplikasi membangun corpus ini merupakan hal pertama yang harus pertama kali dilakukan sebelum melakukan proses pengumpulan data apabila data yang akan diproses belum ada. Tampilan running proses seting crawler dari gambar 8, merupakan hasil dari implementasi dari perancangan sistem diagram alir data gambar 6 pada proses seting crawler.



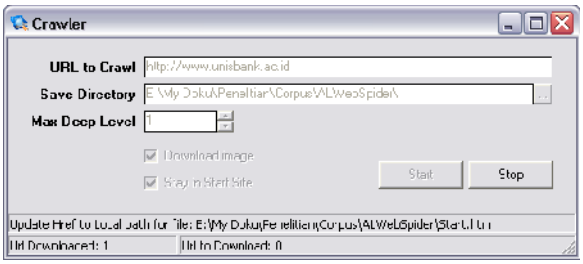
Gambar 8. Tampilan Proses Running Seting Crawler

Proses yang dapat dilakukan pada seting crawler dapat dijelaskan sebagai berikut : pertama, memasukkan isian berupa alamat website yang akan di download datanya, dalam hal ini contoh yang digunakan yaitu alamat web dari unisbank http://www.unisbank.ac.id. Kedua, memasukkan isian alamat direktori untuk menyimpan file hasil download dari mesin crawler ke dalam hardisk, dalam hal ini alamat direktori yang digunakan yaitu e:\my doku\penelitian\corpus\ alwebspider\. Langkah ketiga, memasukkan isian level link yang berfungsi seberapa dalam link yang akan didownload, dalam hal ini contoh yang digunakan adalah nilai satu. Keempat, melakukan check pada isian checkbox download image yang berfungsi untuk menunjukkan bahwa file yang image juga yang termasuk didownload, tidak hanya file teks saja. Kelima,

melakukan check pada checkbox stay in start site dengan maksud bahwa server yang didownload hanya berasal dari www.unisbank.ac.id saja, tidak dari server lain walaupun mungkin saja level link dapat menunjukkan server yang berbeda. Setelah semua isian dilakukan maka berikutnya tinggal menekan tombol start, sehingga proses crawler akan dilakukan.

7. Implementasi Pengumpulan Data menggunakan Crawler

Tampilan running proses pengumpulan data gambar 9 merupakan hasil implementasi dari perancangan sistem diagram alir data gambar 6 pada proses crawling.



Gambar 9. Tampilan Proses Running Crawling

Proses yang dilakukan pada pengumpulan data dengan menggunakan crawler ini merupakan proses untuk mengumpulkan data yang berdasarkan seting yang telah ditentukan sebelumnya, mulai dari alamat web, ditektori tempat menyimpan file hingga kedalaman link download.

8. Implementasi Pencarian, Convert dan Simpan Data

Dalam implementasi ini pada gambar 10 terdapat tiga proses utama yang dilakukan, yang pertama proses pencarian, berikutnya proses convert dan yang terakhir proses simpan data. Proses pencarian dilakukan untuk mencari file data yang telah tersimpan dalam direktori atau drive. File data yang dicari dapat difilter dengan menggunakan menentukan kriteria yang diinginkan misalnya mencari file data dengan format file berekstensi *.htm, maka semua file yang berekstensi htm apabila terdapat dalam direktori yang dicari akan ditampilkan hasilnya. Hasil yang ditampilkan diurutkan berdasarkan urutan pertama kali file yang ditemukan, dalam arti bahwa file yang ditemukan pertama kali akan ditempatkan pada urutan pertama dan sebaliknya file yang ditemukan terakhir akan ditempatkan pada urutan terakhir. Dan dalam melakukan pencarian dari dalam direktori termasuk juga subdirektorinya berdasarkan seting awal apakah file yang dicari hanya pada direktori utama saja atau sekaligus subdirektorinya. Jumlah file yang dapat ditemukan akan ditampilkan hasilnya secara integer misalnya jumlah file yang ditemukan jumlahnya delalpan maka informasi jumlah file ini akan diinformasikan ke user.

Setelah proses pencarian dilakukan makan proses berikutnya yaitu proses converting atau konversi. Maksud dari proses ini yaitu mengkonversi dari semua file dokumen teks dalam format apapun akan dijadikan format teks atau txt. Sedangkan untuk konversi file image maka semua file image yang dihasilkan dari proses pencarian akan dikonversi menjadi file image yang berformat bmp. Tampilan yang dalam implementasi ini akan memperlihatkan semua file dokumen teks dan image secara berurutan berdasarkan urutan dari hasil pencarian akan dikonversi dan tampilannya dapat dilihat oleh user. Selain proses konversi secara bersamaan file-file yang dikonversi tersebut juga akan ditampilkan file-file tersebut letaknya pada direktori mana akan ditampilkan secara tepat, hal ini dilakukan untuk menentukan nantinya pada saat dimasukkan ke dalam database file tersebut dapat diketahui letak persisnya file tersebut berada secara fisik.

Berikutnya setelah proses konversi selesai maka akan dilanjutkan dengan proses penyimpanan data. Proses penyimpanan data dilakukan dari hasil konversi yang telah dilakukan, sebagai contoh sebuah file dokumen teks dengan ekstensi htm dikonversi ke dalam format txt, maka setelah dikonversi ke dalam txt selanjutnya dimasukkan atau disimpan ke dalam database corpus. Saat yang bersamaan pada waktu penyimpanan juga dilakukan penyimpanan terhadap kode dokumen dan alamat lengkap dari letak fisik dokumen tersebut. Dengan begitu setiap data yang tersimpan, baik dokumen teks maupun image akan diketahui letak fisiknya secara tepat.

Gambar 10. Implementasi Pencarian, Converting dan Simpan Data

Sehingga apabila dilakukan pencarian dengan menggunakan basisdata akan lebih mudah dan cepat dari pada melakukan pencarian didalam direktori atau drive.

Apabila ketiga proses tersebut telah dijalankan maka corpus yang dibangun dengan menggunakan aplikasi ini juga telah dilakukan dalam arti bahwa corpus telah selesai dibuat tinggal memanfaatkan untuk digunakan dalam aplikasi-aplikasi yang lain.

HASIL PENELITIAN DAN PEMBAHASAN

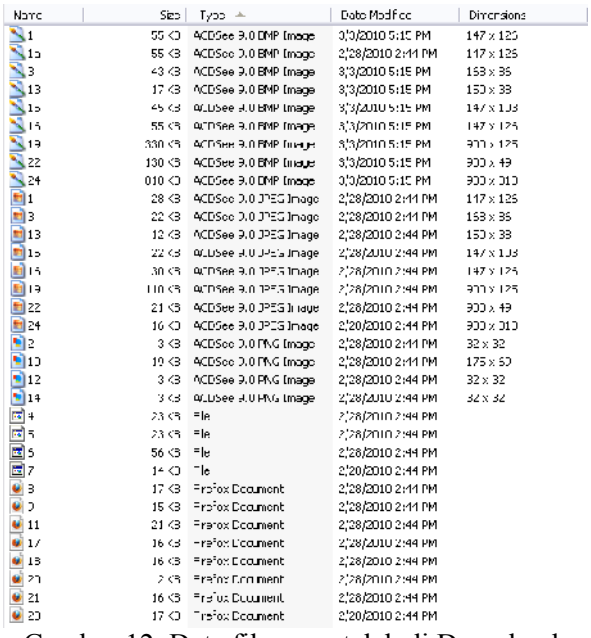
1. Pengumpulan Data dari Alamat Web

Dalam ujicoba ini menggunakan data dari sebuah alamat web yaitu www.unisbank.ac.id (gambar 11) yang kemudian hasil dari download disimpan pada E:\My Doku\Penelitian\Corpus\ALWebSpider\1000, berikutnya dilakukan pencarian file data dari direktori yang telah ditentukan berdasarkan criteria *.htm dan *.jpg yang kemudian hasilnya disimpan dalam database corpus dengan dua buah tabel yaitu tabel image dan tabel teks.

Tabel 11. Halaman Website tempat Data yang akan di Download

2. Hasil Crawling Teks dan Image

Dari mesin crawler yang mengambil data dari alamat web www.unisbank.ac.id mengambil semua file yang ada dengan semua kategori baik file teks, gambar dan sebagainya. Dan hasil download kemudian disimpan dalam sebuah direktori. Berikut hasil dari hasil download yang dapat dilihat pada gambar 12.



Gambar 12. Data file yang telah di Download menggunakan Crawler

3. Tabel Teks dari Hasil Crawling

Dari proses pencarian dari sebuah direktori yang telah ditentukan dan kemudian hasil dari pencarian tersebut dikonversi maka hasilnya akan disimpan dalam database corpus tabel teks. Hasil penyimpanan teks tersebut dapat dilihat hasilnya pada tabel 3.

Tabel 3. Tabel Teks dari Hasil Crawling

id	url	isi
1	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\11.h...	[BLOB - 4.7 KiB]
2	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\17.h...	[BLOB - 2.8 KiB]
3	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\18.h...	[BLOB - 2.4 KiB]
4	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\20.h...	[BLOB - 553 B]
5	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\21.h...	[BLOB - 2.5 KiB]
6	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\23.h...	[BLOB - 2.8 KiB]
7	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\8.ht...	[BLOB - 3.2 KiB]
8	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\9.ht...	[BLOB - 1.8 KiB]

4. Tabel Image dari Hasil Crawling

Dari proses pencarian dari sebuah direktori yang telah ditentukan dan kemudian hasil dari pencarian tersebut dikonversi maka hasilnya akan disimpan dalam database corpus tabel teks. Hasil penyimpanan teks tersebut dapat dilihat hasilnya pada tabel 4.

Tabel 4. Tabel Image dari Hasil Crawling

id	url	isi
1	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\1.jp...	[BLOB - 54.7 KiB]
2	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\13.j...	[BLOB - 16.8 KiB]
3	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\15.j...	[BLOB - 44.7 KiB]
4	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\16.j...	[BLOB - 54.7 KiB]
5	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\19.j...	[BLOB - 329.6 KiB]
6	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\22.j...	[BLOB - 129.3 KiB]
7	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\24.j...	[BLOB - 817.4 KiB]
8	E:\My Doku\Penelitian\Corpus\ALWebSpider\1000\3.jp...	[BLOB - 42.4 KiB]

KESIMPULAN

1. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan pada sistem maka dapat disimpulkan bahwa :

a. Berdasarkan hasil uji coba aplikasi yang dilakukan dari proses seting crawler dengan alamat web www.unisbank.ac.id, dengan folder untuk penyimpanan data hasil download pada E:\My Doku\Penelitian\Corpus\ALWebSpider\1000, dengan max deep level pada nilai satu, image ikut di download, dan proses donwload pada server yang sama, maka menghasilkan tiga puluh tiga file baik file dokumen teks maupun image.

b. Aplikasi mampu melakukan download dari sebuah alamat web dengan otomatis dengan ketentuan dapat dilakukan oleh user. Semua file yang dapat didownload akan diambil semuanya tanpa terkecuali. Sehingga akan mempermudah user dalam pengumpulan data tanpa harus mendownload satu-per-satu file.

c. Aplikasi mampu menampilkan hasil pencarian dokumen dan mengurutkannya berdasarkan urutan dari penemuan dari file data yang dicari, dalam arti dokumen data yang ditemukan pertama kali akan ditempatkan diurutan pertama sedangkan dokumen data yang ditemukan terakhir akan ditempatkan pada urutan paling bawah. Dalam uji implementasi dapat dilihat bahwa urutan pertama adalah file 1.bmp dan yang terakhir adalah 9.htm

d. Aplikasi mampu melakukan konversi dari dokumen teks dengan berbagai format data ke dalam bentuk dokumen teks txt, juga dalam melakukan konversi pada semua format file image ke dalam bentuk format bmp. Konversi dilakukan untuk menyamakan format untuk dapat mempermudah dalam penyimpanan dalam database.

e. Aplikasi mampu menyimpan dokumen teks dan image dalam tabel teks dan tabel image secara otomatis dari semua hasil pencarian dan konversi yang telah dilakukan pada proses sebelumnya. Sehingga mempermudah user apabila mempunyai data yang besar tanpa harus menginput satu-per-satu file data ke dalam database.

2. Saran

a. Aplikasi ini dalam menyimpan file data ke dalam database disamakan format filenya sehingga dibutuhkan proses konversi, sehingga perlu dikembangkan lagi tanpa ada proses konversi yang diharapkan akan mempercepat waktu proses secara keseluruhan dengan menyimpan data dalam berbagai format.

b. Aplikasi ini perlu dikembangkan dan di ujicoba dengan lebih komprehensif dalam

berbagai format data yang lain baik dalam bentuk teks maupun image. Sehingga aplikasi ini dapat digunakan secara universal oleh pengguna dengan berbagai macam data.

DAFTAR PUSTAKA

Dennis De Champeaux, Dauglas Lea and Penelope Faure, *Object Oriented System Development,* Addison Wesley Publishing Company, California, 1994

Grady Booch, Object - Oriented Analysis And Design, The Benjamin / Cummings Publishing Company, Inc, California, 1994

Lou Burnard L., 1996, *Using SGML for Linguistic Analysis: the case of the BNC,* http://users.ox.ac.uk/~lou/wip/Boston/

Marco Cantu, 2007, Mastering Delphi 7, Sybex

McEnery, T. and Wilson, A., 2001, *Corpus Linguistics 2nd Edition*. Edinburgh University Press

Pressman, R.S. 1997, *Software engineering : a practitioner's approach*, McGraw-Hill, New York

Salton, G. 1989, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley

Sheridan, P. and Ballerini J.P., 1996, ”Experiments in Multilingual Information Retrieval using the SPIDER System”, *In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*

Yang, Y., dan Wilbur, J., 1996, *Using corpus statistics to remove redundant words in text categorization,* JASIS